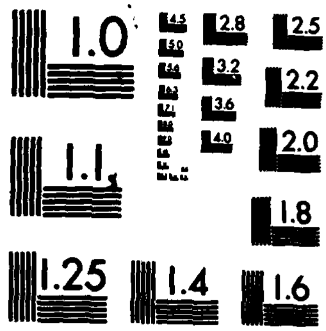
AD-A129 170 A NOTE ON THE GEOMETRY OF KULLBACK-LEIBLER INFORMATION NUMBERS(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER W LOH APR 83 MRC-TSR-2506 DAAG29-80-C-0041 1/1

UNCLASSIFIED F/G 12/1 NL

END
DATE
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(2)

MRC Technical Summary Report #2506

A NOTE ON THE GEOMETRY OF KULLBACK-LEIBLER INFORMATION NUMBERS

Wei-Yin Loh

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

April 1983

(Received March 23, 1983)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

A NOTE ON THE GEOMETRY OF KULLBACK-LEIBLER INFORMATION NUMBERS

Wei-Yin Loh

Technical Summary Report #2506

April 1983

ABSTRACT

Csiszar (1975) has shown that Kullback-Leibler information numbers possess some geometrical properties much like those in Euclidean geometry. This paper extends these results by characterizing the shortest line between two distributions as well as the midpoint of the line. It turns out that the distributions comprising the line have applications to the problem of testing separate families of hypotheses.

AMS (MOS) Subject Classifications: Primary - 60-00-E05, 62B10
Secondary - 62F03

Key Words: Kullback-Leibler information, geometry of probability distributions, minimax, embedding

Work Unit Number 4 - Statistics and Probability

Department of Statistics and Mathematics Research Center, University of Wisconsin, Madison, WI 53705.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2 and Nos. MCS-7825301 and MCS-7903716.

SIGNIFICANCE AND EXPLANATION

The Kullback-Leibler information number is a well-known measure of statistical distance between probability distributions. Previous authors have shown that when endowed with this distance measure, the space of probability distributions possesses geometrical properties analogous to Euclidean geometry. This paper proves a new geometrical property by showing that one can in fact define the shortest line between two probability distributions as well as its mid-point.

It turns out that the probability distributions comprising this line have long ago been used as a tool in the important problem of testing statistical hypotheses involving nuisance parameters. Apart from pure mathematical convenience, there has been little justification for its use. The results in this paper are the first attempt at such explanation.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

A NOTE ON THE GEOMETRY OF KULLBACK-LEIBLER INFORMATION NUMBERS

Wei-Yin Loh

1. Introduction

Csiszar (1975) has shown that if we use the Kullback-Leibler information number as a measure of distance between (probability) distributions, certain analogies exist between the properties of distributions and Euclidean geometry. In particular, he proved an analogue of Pythagoras' theorem. In this note we extend these geometrical properties by defining the "shortest line" between two distributions and the "mid-point" of the line. It turns out that the distributions comprising such a line are precisely those whose densities are exponential linear combinations of the densities of the two distributions at the end-points.

The idea of taking exponential linear combinations of densities is not new. For example, it appears in Cox (1961), Atkinson (1970) and Brown (1971) as a mathematically convenient means of embedding two families of distributions into a larger family. Our results in section 4 show that in fact there is a deeper mathematical property behind this choice of embedding, namely that the distributions in the embedding are really those distributions that are closest (in the Kullback-Leibler sense) to the two original families.

Department of Statistics and Mathematics Research Center, University of Wisconsin, Madison, WI 53705.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2 and Nos. MCS-7825301 and MCS-7903716.

2. Notations and definitions

Recall that if F and G are two distributions on the same measurable space, the Kullback-Leibler information number $K(F,G)$ is defined as

$$K(F,G) = \begin{cases} \int \log(dF/dG)dF, & \text{if } F << G \\ +\infty & , \text{otherwise} \end{cases}$$

where "$F << G$" means that F is absolutely continuous with respect to G. It is well known that $K(F,G)$ is well-defined, nonnegative, and is equal to zero if and only if $F(B) = G(B)$ for all measurable sets B.

We need the following definitions in the rest of this paper.

Definition 2.1. A distribution P is closer to F and G than Q is if

$$K(P,F) \leq K(Q,F) \quad \text{and} \quad K(P,G) \leq K(Q,G)$$

with at least one inequality being strict. In symbols we write $P \underset{FG}{<} Q$ (or $P < Q$ if it is clear from the context what F and G are).

Definition 2.2. P is a mid-point of F and G if $K(P,F) = K(P,G)$ and there does not exist Q for which $Q \underset{FG}{<} P$.

Definition 2.3. P is minimax for F and G if $\max(K(P,F), K(P,G)) = \min_Q\{\max(K(Q,F), K(Q,G))\}$ where the min is taken over the space of all distributions.

Throughout this paper, μ denotes a measure that dominates both F and G; and $f(x)$, $g(x)$ are their respective densities relative to μ. For convenience, we let A denote the set

$$A = \{x : f(x)g(x) > 0\} \quad , \tag{2.1}$$

and let $P_\lambda (0 < \lambda < 1)$ be the distribution with density (with respect to μ) given by

$$p_\lambda(x) = \begin{cases} k_\lambda g^\lambda(x) f^{1-\lambda}(x) & \text{on } A \\ 0 & \text{otherwise} \end{cases} \quad , \tag{2.2}$$

where $k_\lambda^{-1} = \int_A g^\lambda f^{1-\lambda} d\mu$, and

$$P = \{P_\lambda,\ 0 < \lambda < 1\} \cup \{F,G\} \quad . \tag{2.3}$$

(Note that if F and G are mutually absolutely continuous, $P_0 = F$, $P_1 = G$ and P is an exponential family.) Finally we need the function

$$J(\lambda) = \int_A g^\lambda f^{1-\lambda} \log(g/f) d\mu,\ 0 < \lambda < 1 \quad . \tag{2.4}$$

We will often abbreviate $K(P_\lambda,\cdot)$ to $K(\lambda,\cdot)$.

3. Preliminary lemmas.

We will assume throughout that μ is a σ-finite measure and F and G are two (fixed) distributions, not necessarily mutually absolutely continuous.

Lemma 3.1. Suppose that $\mu(A) > 0$. Then k_λ and $J(\lambda)$ are both differentiable in $(0,1)$ and continuous at $\lambda = 0, 1$ (with $J(0)$ and $J(1)$ possibly infinite).

Proof. Since $k_\lambda^{-1} = \int_A \exp(\lambda \log(g/f))f\, d\mu$ and $J(\lambda)$ is its first derivative, differentiability in $(0,1)$ follows from a well-known result on integrals of exponential densities (see e.g. Lehmann (1959)). To see that the functions are continuous at the end-points, split A into the sets $A(f) = A \cap \{f > g\}$ and $A(g) = A \cap \{f < g\}$, and use dominated convergence to obtain the result for k_λ. To prove the same for $J(\lambda)$, first observe that nonnegativity of $K(G,F)$ implies that

$$0 < \int_A [\log(g/f)]^- g\, d\mu < \infty \quad .$$

Therefore we may take limits as $\lambda \downarrow 0$ in the inequality

$$\int_A [\log(g/f)]^- g^\lambda f^{1-\lambda} d\mu < \{\int_A [\log(g/f)]^- g\, d\mu\}^\lambda \{\int_A [\log(g/f)]^- f\, d\mu\}^{1-\lambda}$$

to obtain

$$(3.1) \qquad \lim_{\lambda \downarrow 0} \int_A [\log(g/f)]^- g^\lambda f^{1-\lambda} d\mu < \int_A [\log(g/f)]^- f\, d\mu \quad .$$

Fatou's lemma shows that the reverse inequality holds, so in fact exact equality obtains in (3.1). Now by monotone convergence

$$\lim_{\lambda \downarrow 0} \int_A [\log(g/f)]^+ g^\lambda f^{1-\lambda} d\mu = \int_A [\log(g/f)]^+ f\, d\mu \quad .$$

This proves that $J(\lambda)$ is continuous at $\lambda = 0$. A similar argument does it for $\lambda = 1$.

Lemma 3.2. As functions of λ, both $K(\lambda,F)$ and $K(\lambda,G)$ are differentiable in $(0,1)$ and continous at $\lambda = 0, 1$. $K(\lambda,F)$ is non-decreasing and $K(\lambda,G)$ is non-increasing in $[0,1]$.

Proof. The first assertion follows from the preceding lemma and the relations

$$K(\lambda,F) = \log k_\lambda + \lambda k_\lambda J(\lambda) \tag{3.2}$$

$$K(\lambda,G) = \log k_\lambda - (1-\lambda)k_\lambda J(\lambda) \quad . \tag{3.3}$$

Differentiation yields, for $0 < \lambda < 1$,

$$\begin{aligned} \lambda^{-1}(d/d\lambda)K(\lambda,F) &= -(1-\lambda)^{-1}(d/d\lambda)K(\lambda,G) \\ &= \mathrm{Var}_\lambda\{\log(g(X)/f(X))\} \geq 0 \quad . \end{aligned} \tag{3.4}$$

This proves the second assertion. It is easy to see that strict inequality holds in (3.4) for some $0 < \lambda < 1$ if and only if it holds for all $0 < \lambda < 1$.

Lemma 3.3. Suppose that $\mu(A) > 0$. Let Q be such that $K(Q,F)$ and $K(Q,G)$ are both finite and define

$$\begin{aligned} r(\lambda) &= \int \log(p_\lambda/f)dQ \\ s(\lambda) &= \int \log(p_\lambda/g)dQ \quad . \end{aligned} \tag{3.5}$$

Then (i) $r(\lambda)$ and $s(\lambda)$ are finite and continuous in $[0,1]$, and (ii) if for some $0 < \lambda < 1$,

$$r(\lambda) = K(\lambda,F) \tag{3.6}$$

then $s(\lambda) = K(\lambda,G)$.

Proof. The finiteness of $K(Q,F)$ and $K(Q,G)$ means that Q is absolutely continuous with respect to P_λ for all λ in $[0,1]$. Therefore we may write

$$r(\lambda) = \log k_\lambda + \lambda(K(Q,F) - K(Q,G))$$

$$s(\lambda) = \log k_\lambda - (1-\lambda)(K(Q,F) - K(Q,G)) \quad .$$

Assertion (i) now follows from Lemma 3.1. To get (ii) use the fact that

$$K(\lambda,F) = \log k_\lambda + \lambda(K(\lambda,F) - K(\lambda,G))$$

and $K(\lambda,G) = \log k_\lambda - (1-\lambda)(K(\lambda,F) - K(\lambda,G))$.

The proof of the next lemma is trivial. A more general version appears in Csiszar (1975).

<u>Lemma 3.4</u>. Let P, Q, R be three distinct distributions such that $P << R$ and $K(Q,P) < \infty$. Then

$$\int \log(dP/dR)dQ = K(P,R)$$

if and only if

$$K(Q,R) = K(Q,P) + K(P,R) \quad .$$

A similar result holds if both "=" signs are replaced with ">" signs.

4. Main results

We now prove our main theorem, which says that the exponential embedding $\mathcal{P}$ in (2.3) is in some sense "complete".

Theorem 4.1. For any Q not belonging to $\mathcal{P}$, there is P in $\mathcal{P}$ such that $P \underset{FG}{\leq} Q$.

Proof. The result is easy if F and G are mutually singular since then $K(F,G) = K(G,F) = \infty$ and we may take $P = F$ if $K(Q,G) = \infty$ and $P = G$ otherwise. So suppose $\mu(\lambda) > 0$, and without loss of generality further assume that both $K(Q,F)$ and $K(Q,G)$ are finite. Then $Q \ll P_\lambda$ for all $0 < \lambda < 1$. Let r and s be defined as in (3.5). By Lemmas 3.2 and 3.3, $K(\lambda,F)$ and $r(\lambda)$ are continuous functions of λ in $[0,1]$. We consider three cases according to whether these two graphs intersect.

(I). Suppose $r(\lambda) = K(\lambda,F)$ for some $0 < \lambda < 1$. Then

$$K(Q,\lambda) = \lambda K(Q,G) + (1-\lambda)K(Q,F) - \log k_\lambda < \infty \tag{4.1}$$

and Lemma 3.4 implies that

$$K(Q,F) = K(Q,\lambda) + K(\lambda,F) > K(\lambda,F) \quad .$$

Further, by Lemma 3.3, $s(\lambda) = K(\lambda,G)$. Reversing the roles of r and s, and F and G, we also get $K(Q,G) > K(\lambda,G)$. Hence $P_\lambda \underset{FG}{\leq} Q$.

(II). Suppose $r(\lambda) > K(\lambda,F)$ for all $0 < \lambda < 1$. Continuity yields $r(1) > K(1,F)$ and since $K(Q,1) < \infty$ by (4.1), we can use Lemma 3.4 to deduce that $K(Q,F) > K(Q,1) + K(1,F) > K(1,F)$. Since $K(Q,G) = K(Q,1) + K(1,G) > K(1,G)$, it follows that $P_1 \underset{FG}{\leq} Q$.

(III). The case $r(\lambda) < K(\lambda,F)$ for all $0 < \lambda < 1$ is similar to (II).

According to Definition 2.2, the above theorem implies that the mid-point M of F and G belongs to $\mathcal{P}$ whenever the former exists. The following corollaries give conditions for the existence of M.

Corollary 4.1. (i) If F and G are mutually absolutely continuous, M exists and equals P_λ for some unique λ in $(0,1)$. (ii) If F and G are mutually singular, M does not exist. (iii) M is unique whenever it exists.

Proof. Assertion (i) follows from the fact that if F and G are mutually absolutely continuous and distinct from each other, then $K(0,F) = K(1,G) = 0$, and both $K(\lambda,F)$ and $K(\lambda,G)$ are strictly monotone for $0 < \lambda < 1$. Assertion (ii) is immediate from Theorem 4.1 since $P = \{F,G\}$ if F and G are mutually singular. To prove assertion (iii), suppose that F and G are not mutually singular and M exists. If there are $\lambda_1 \neq \lambda_2$ in $[0,1]$ such that P_{λ_1} and P_{λ_2} are both mid-points of F and G, then

$$K(\lambda_1,F) = K(\lambda_1,G) = K(\lambda_2,F) = K(\lambda_2,G)$$

and it follows from (3.4) that $g(x)/f(x)$ is constant a.e. (μ) on A. This implies that $P_\lambda = P_0$ for all $0 < \lambda < 1$ and hence that M is unique.

Corollary 4.2. Suppose F and G are not mutually singular. Then the mid-point M *exists if and only if*

$$J(\lambda) = 0 \text{ for some } 0 < \lambda < 1 \; , \tag{4.2}$$

in which case $M = P_\lambda$.

Proof. According to Theorem 4.1, M exists if and only if

$$K(\lambda,F) = K(\lambda,G) < \infty \text{ for some } 0 < \lambda < 1 \; . \tag{4.3}$$

It is clear from (3.2) and (3.3) that this is equivalent to (4.2).

Corollary 4.1 states that mutual singularity of F and G is a sufficient condition for the non-existence of the mid-point. The following example shows that the condition is not necessary.

Example 4.1. Let F be the uniform distribution on $(0,3)$ and G be uniform on $(1,2)$. Then $P_\lambda = G$ for all $0 < \lambda < 1$ and (4.3) does not hold for any λ. There is thus no mid-point.

The P_λ (or G) in this example is "minimax" according to Definition 2.3. It turns out that minimax distributions exist always. Uniqueness may be lost but only in trivial cases. This is made explicit in the next corollary.

Corollary 4.3. (i) A minimax distribution always exists. (ii) If F and G are not mutually singular, the minimax distribution is unique. (iii) If F and G are mutually singular, every distribution is minimax. (iv) Every mid-point is unique minimax.

Proof. Since every mid-point is minimax by definition, assertion (iv) is immediate from Corollary 4.1. It remains to prove assertions (i) - (iii) only for the case when the mid-point does not exist. To prove assertion (ii), suppose that F and G are not mutually singular. It is clear from (4.3) that the mid-point does not exist if and only if the graphs of $K(\lambda,F)$ and $K(\lambda,G)$ fail to intersect in $[0,1]$. But from (3.4) either

$$(4.4)\qquad (d/d\lambda)K(\lambda,F) > 0 \quad\text{and}\quad (d/d\lambda)K(\lambda,G) < 0 \quad\text{for all}\quad 0 < \lambda < 1$$

or

$$(4.5)\qquad (d/d\lambda)K(\lambda,F) = (d/d\lambda)K(\lambda,G) = 0 \quad\text{for all}\quad 0 < \lambda < 1 \quad .$$

Therefore either $K(0,F) > K(0,G)$ or $K(1,F) < K(1,G)$. Assume, without loss of generality, that $K(0,F) > K(0,G)$. First suppose that (4.4) holds. Then for all $0 < \lambda < 1$

$$(4.6)\qquad \max(K(0,F),\ K(0,G)) < \max(K(\lambda,F),\ K(\lambda,G)) \quad .$$

If $K(G,F) < \infty$, then $G << F$, $P_1 = G$ and (4.6) yields

$$(4.7)\qquad \max(K(0,F),\ K(0,G)) < \max(K(G,F),\ K(G,G)) \quad .$$

Clearly (4.7) is trivially true also if $K(G,F) = \infty$. A similar argument shows that

$$(4.8)\qquad \max(K(0,F),\ K(0,G)) \leq \max(K(F,F),\ K(F,G))$$

with equality if and only if $P_0 = F$. Now (4.6) - (4.8) shows that P_0 uniquely minimizes $\max(K(P,F),\ K(P,G))$ over all $P \in \mathcal{P}$. We conclude from

Theorem 4.1 that P_0 is unique minimax for F and G. If instead (4.5) obtains, then as the proof of Corollary 4.1 shows, $P_\lambda = P_0$ for all $0 < \lambda < 1$. Further $K(0,F) = K(0,G) = 0$. Since (4.7) and (4.8) are trivially satisfied, P_0 is again unique minimax. This completes the proof of assertion (ii). Assertion (iii) follows from observing that if F and G are mutually singular, then at least one of $K(Q,F)$ and $K(Q,G)$ is infinite for any distribution Q. Assertion (i) is a consequence of (ii) - (iv).

5. Example.

We end this discussion with two examples.

Example 5.1 (Binomial). Let F be $Bin(n,p_1)$ (binomial with n trials and success probability p_1) and G be $Bin(n,p_2)$. Write $q_i = 1-p_i$. Then every member in $\mathcal{P}$ is binomial and the mid-point M is $Bin(n,p)$ where $p = \log(q_2/q_1)/\log(p_1q_2/p_2q_1)$. This formula applies and yields p between p_1 and p_2 only when neither p_1 nor p_2 is 0 or 1. If $p_1 = 0$ and $0 < p_2 < 1$ for example, the formula gives $p = 0$. The reason for this strange result is that here there is no mid-point since $\mathcal{P} = \{F\}$. It can be shown that if both p_1 and p_2 are neither 0 nor 1, then p lies strictly between the two p's. In the special case that $p_1 = 1-p_2$, then $p = \frac{1}{2}$ as expected. The formula for p suggests a new way of "scaling" the binomial family.

Example 5.2 (Normal). Let F be $N(\theta_1,\sigma_1^2)$ (normal with mean θ_1 and variance σ_1^2) and G be $N(\theta_2,\sigma_2^2)$. Then the members of $\mathcal{P}$ are also normal distributions. If $\sigma_1 = \sigma_2$, M is $N(\frac{1}{2}(\theta_1+\theta_2), \sigma_1^2)$; and if $\theta_1 = \theta_2$, $\sigma_1 \neq \sigma_2$, then M is $N(\theta_1,\sigma^2)$ where

$$\sigma^2 = \sigma_1^2\sigma_2^2 \log(\sigma_2^2/\sigma_1^2)/(\sigma_2^2 - \sigma_1^2) \quad .$$

It can be verified that σ always lies between σ_1 and σ_2.

Acknowledgement. The author is grateful to E. L. Lehmann for many helpful comments.

REFERENCES

[1] Atkinson, A. C. (1970). A method for discriminating between models. J. Roy. Statist. Soc. (B) 32, 323-353.

[2] Brown, L. D. (1971). Non-local asymptotic optimality of appropriate likelihood ratio tests. Ann. Math. Statist. 42, 1206-1240.

[3] Cox, D. R. (1961). Tests of separate families of hypotheses. Proc. Fourth Berkeley Symp. 1, 105-123.

[4] Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. Ann. Probability, 3, 146-158.

[5] Lehmann, E. L. (1959). Testing Statistical Hypotheses. Wiley, New York.

WYL/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2506	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A NOTE ON THE GEOMETRY OF KULLBACK-LEIBLER INFORMATION NUMBERS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Wei-Yin Loh		8. CONTRACT OR GRANT NUMBER(s) MCS-7927062, Mod. 2; DAAG29-80-C-0041; MCS-7825301; MCS-7903716
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS See Item 18 below		12. REPORT DATE April 1983
		13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Kullback-Leibler information, geometry of probability distributions, minimax, embedding

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Csiszar (1975) has shown that Kullback-Leibler information numbers possess some geometrical properties much like those in Euclidean geometry. This paper extends these results by characterizing the shortest line between two distributions as well as the midpoint of the line. It turns out that the distributions comprising the line have applications to the problem of testing separate families of hypotheses.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FILM

7